US006714925B1

(12) **United States Patent**
**Barnhill et al.**

(10) **Patent No.: US 6,714,925 B1**
(45) **Date of Patent: *Mar. 30, 2004**

(54) **SYSTEM FOR IDENTIFYING PATTERNS IN BIOLOGICAL DATA USING A DISTRIBUTED NETWORK**

(75) Inventors: **Stephen Barnhill**, Savannah, GA (US); **Isabelle Guyon**, Berkeley, CA (US); **Jason Weston**, New York, NY (US)

(73) Assignee: **Barnhill Technologies, LLC**, Savannah, GA (US)

( * ) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 260 days.

This patent is subject to a terminal disclaimer.

(21) Appl. No.: **09/633,627**

(22) Filed: **Aug. 7, 2000**

**Related U.S. Application Data**

(63) Continuation-in-part of application No. 09/578,011, filed on May 24, 2000, now Pat. No. 6,658,395, and a continuation-in-part of application No. 09/568,301, filed on May 9, 2000, and a continuation-in-part of application No. 09/303,386, filed on May 1, 1999, and a continuation-in-part of application No. 09/303,387, filed on May 1, 1999, now Pat. No. 6,128,608, and a continuation-in-part of application No. 09/303,389, filed on May 1, 1999, now abandoned, and a continuation-in-part of application No. 09/305,345, filed on May 1, 1999, now Pat. No. 6,157,921.

(60) Provisional application No. 60/191,219, filed on Mar. 22, 2000, provisional application No. 60/184,596, filed on Feb. 24, 2000, provisional application No. 60/168,703, filed on Dec. 2, 1999, and provisional application No. 60/161,806, filed on Oct. 27, 1999.

(51) **Int. Cl.**[7] ..... **G06F 17/00**

(52) **U.S. Cl.** ..... **706/48**; 706/16

(58) **Field of Search** ..... 706/16, 48

(56) **References Cited**

U.S. PATENT DOCUMENTS

| | | | | |
|---|---|---|---|---|
| 4,881,178 | A | 11/1989 | Holland et al. | 706/12 |
| 5,138,694 | A | 8/1992 | Hamilton et al. | 706/52 |
| 5,649,068 | A | 7/1997 | Boser et al. | 706/12 |
| 5,809,144 | A | 9/1998 | Sirbu et al. | 705/26 |
| 5,950,146 | A | 9/1999 | Vapnik | 702/153 |
| 6,128,608 | A | 10/2000 | Barnhill | 706/16 |
| 6,157,921 | A * | 12/2000 | Barnhill | 706/16 |
| 6,282,523 | B1 * | 8/2001 | Tedesco et al. | 705/45 |
| 6,427,141 | B1 * | 7/2002 | Barnhill | 706/16 |

OTHER PUBLICATIONS

Yan, Yonghong et al., "Experiments for an approach to language identification with conversational telephone speech" 1995 IEEE International Conference, pp. 789–792.

Osuna, Edgar et al., "An Improved Training Algorithm for Support Vector Machines", 1997 IEEE International Conference, pp. 276–285.

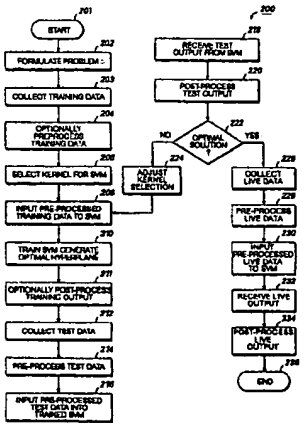
(List continued on next page.)

*Primary Examiner*—George B. Davis
(74) *Attorney, Agent, or Firm*—Kilpatrick Stockton LLP

(57) **ABSTRACT**

System for enhancing knowledge discovery from data using a learning machine in general and a support vector machine in particular in a distributed network environment. A customer transmits training data, test data and live data to a vendor's server from a remote source, via a distributed network. The training biological data, test biological data and live biological data is stored in a storage device. Training biological data is then pre-processed in order to add meaning thereto. Pre-processing data involves transforming the biological data points and/or expanding the biological data points. Live biological data is pre-processed and input into the trained and tested learning machine. The live output from the learning machine is then post-processed into a computationally derived alphanumerical classifier for interpretation by a human or computer automated process.

**16 Claims, 33 Drawing Sheets**



US-PAT-NO: 6714925

DOCUMENT-IDENTIFIER: US 6714925 B1

TITLE: System for identifying patterns in biological data using a distributed network

DATE-ISSUED: March 30, 2004


INVENTOR-INFORMATION:

| NAME | CITY | STATE | ZIP CODE | COUNTRY |
|---|---|---|---|---|
| Barnhill; Stephen | Savannah | GA | N/A | N/A |
| Guyon; Isabelle | Berkeley | CA | N/A | N/A |
| Weston; Jason | New York | NY | N/A | N/A |


US-CL-CURRENT: 706/48, 706/16


ABSTRACT:

System for enhancing knowledge discovery from data using a learning machine in general and a support vector machine in particular in a distributed network environment. A customer transmits training data, test data and live data to a vendor's server from a remote source, via a distributed network. The training biological data, test biological data and live biological data is stored in a storage device. Training biological data is then pre-processed in order to add meaning thereto. Pre-processing data involves transforming the biological data points and/or expanding the biological data points. Live biological data is pre-processed and input into the trained and tested learning machine. The live output from the learning machine is then post-processed into a computationally derived alphanumerical classifier for interpretation by a human or computer automated process.


16 Claims, 54 Drawing figures

Exemplary Claim Number: 1

Number of Drawing Sheets: 33

---------- KWIC ---------

Detailed Description Text - DETX (36):

As mentioned above, the exemplary optimal categorization method 300 may be used in pre-processing data and/or post-processing the output of a learning machine. For example, as a pre-processing transformation step, the exemplary optimal categorization method 300 may be used to extract classification information from **raw data**. As a post-processing technique, the exemplary optimal range categorization method may be used to determine the optimal cut-off values for markers objectively based on data, rather than relying on ad hoc approaches. As should be apparent, the exemplary optimal categorization method 300 has applications in pattern recognition, classification, regression problems, etc. The exemplary optimal categorization method 300 may also be used as a stand-alone categorization technique, independent from SVMs and other learning machines. An exemplary stand-alone application of the optimal categorization method 300 will be described with reference to FIG. 8.

Detailed Description Text - DETX (57):
Software customization and development allow optimization of activities on the BSVP. Concurrency in sections of SVM processes is exploited in the most advantageous manner through the hybrid parallelization provided by the BSVP hardware. The software implements full cycle support from **raw data** to implemented solution. A database engine provides the storage and flexibility required for pre-processing **raw data**. Custom developed routines automate the pre-processing of the data prior to SVM training. Multiple transformations and data manipulations are performed within the database environment to generate candidate training data.

Detailed Description Text - DETX (118):
A more detailed discussion of the methods of a preferred embodiment follow. An SVM Recursive Feature Elimination (RFE) was run on the **raw data** to assess the validity of the method. The colon cancer data samples were split randomly into 31 examples for training and 31 examples for testing. The RFE method was run to progressively downsize the number of genes by each time dividing it by 2. The preprocessing of the data was that for each gene expression value, the mean was subtracted and then the resultant was divided by the standard deviation.

Detailed Description Text - DETX (125):
The initial preprocessing steps of the data were described by Alon et al. The data was further preprocessed in order to make the data distribution less skewed. FIG. 15 shows the distributions of gene expression values across tissue samples for two random genes (cumulative number of samples of a given expression value) which is compared with a uniform distribution. Each line represents a gene. 15A and B show the **raw data**; 15C and D are the same data after taking the log. By taking the log of the gene expression values the same curves result and the distribution is more uniform. This may be due to the fact that gene expression coefficients are often obtained by computing the ratio of two values. For instance, in a competitive hybridization scheme, DNA from two samples that are labeled differently are hybridized onto the array. One obtains at every point of the array two coefficients corresponding to the fluorescence of the two labels and reflecting the fraction of DNA of either sample that hybridized to the particular gene. Typically, the first initial preprocessing step that is taken is to take the ratio a/b of these two values. Though this initial preprocessing step is adequate, it may not be optimal when the two values are small. Other initial preprocessing steps include (a-b)/(a+b) and (log a-log b)/(log a+log b).

Detailed Description Text - DETX (127):
FIG. 16 shows the distribution of gene expression values across genes for all tissue samples. 16A shows the **raw data** and 16B shows the inv erf. The shape is roughly that of an erf function, indicating that the density follows approximately the Normal law. Indeed, passing the data through the inverse erf function yields almost straight parallel lines. Thus, it is reasonable to normalize the data by subtracting the mean. This preprocessing step is also suggested by Alon et al. This preprocessing step is supported by the fact that there are variations in experimental conditions from microarray to microarray. Although standard deviation seems to remain fairly constant, the other preprocessing step selected was to divide the gene expression values by the standard deviation to obtain centered data of standardized variance.

Detailed Description Text - DETX (150):
**Unsupervised Clustering**

Detailed Description Text - DETX (151):

To overcome the problems in gene ranking alone, the data was preprocessed with an **unsupervised clustering** method. 20 Genes were grouped according to resemblance (with a given metric). Cluster centers are then used instead of genes themselves and processed by SVM RFE. The result was nested subsets of cluster centers. An optimum subset size can be chosen with the same cross-validation method used before. The cluster centers can then be replaced either element of the cluster.

Detailed Description Text - DETX (154):

With **unsupervised clustering,** a set of informative genes is defined, but there is no guarantee that the genes not retained do not carry information. When RFE was used on all QT_clust clusters plus the remaining non-clustered genes (singleton clusters), the performance curves were quite similar, though the top set of gene clusters selected was completely different and included mostly singletons. The genes selected in Table 1 are organized in a structure: within a cluster, genes are redundant, across clusters they are complementary.

Detailed Description Text - DETX (164):

Compared to the **unsupervised clustering** method and results, the supervised clustering method, in this instance, does not give better control over the number of examples per cluster. Therefore, this method is not as good as **unsupervised clustering** if the goal is to be able to select from a variety of genes in each cluster. However, supervised clustering may show specific clusters that have relevance for the specific knowledge being determined. In this particular embodiment, in particular, a very large cluster of genes was found that contained several muscle genes that may be related to tissue composition and may not be relevant to the cancer vs. normal separation. Thus, those genes are good candidates for elimination from consideration as having little bearing on the diagnosis or prognosis for colon cancer.

Detailed Description Text - DETX (244):

Though not wishing to be bound by any particular theory, RFE ranking can be thought of as producing nested subsets of features of increasing size that are optimal in some sense. Individually, a feature that is ranked better than another one may not separate the data better. In fact, there are features with any rank that are highly correlated with the first ranked feature. One way of adding a correlation dimension to the simple linear structure provided by SVM RFE is to cluster genes according to a given correlation coefficient. **Unsupervised clustering** in pre-processing for SVM RFE was shown in the present application. The cluster centers were then used as features to be ranked. Supervised clustering was also used as a post-processing for SVM RFE. Top ranking features were also used as cluster centers. The remaining rejected features were clustered to those centers.